Proceedings of the First PhD Symposium on Sustainable Ultrascale Computing Systems (NESUS PhD 2016)
Timisoara, Romania

Jesus Carretero, Javier Garcia Blas
Dana Petcu
(Editors)

February 8-11, 2016

A Framework for Knowledge Management using Complex Networks Methods

Alex Becheru

University of Craiova, Romania
becheru@gmail.com

Abstract

In a world where complexity is constantly increasing due to the technological advancement, large scale of data available and increased interaction between various phenomena there was a need for a field of study to model and understand such complex systems. One such field of research is called Complex Networks Analysis *(CNA) or Network Science. The heart of this research field leverages on Graph Theory and Computer Science. In this paper we shall briefly present a common framework for knowledge management using CNA methods. The power of the framework shall be proven by extracting knowledge from various heterogeneous domains like: Tourism, E-learning, Freight Transportation , and Organisational Analysis.*

Keywords Complex Networks, Knowledge Management, Graph Theory, Tourism, E-Learning, Organisational Analysis

I. Introduction

Our current understanding of the surrounding world shows us that nature is formed out of complex interconnecting systems. Networks created by these systems support phenomena that are far from being deterministic trough traditional methods. Each element influences the network, while the network puts its mark on every element. Now we can say with certainty that the *butterfly effect* imagined by Edward Lorenz is truly possible.

The complexity of real world networks comes from the modelling and evaluation of overlapping and interdependent phenomena, that are neither purely regular nor purely random. Also complexity may come with the sheer size of the network itself.

In order to understand complex interconnected systems a new field of research emerged – *Network Science* (NS) or *Complex Networks Analysis* (CNA). The heart of this new research field leverages on *Graph Theory* and *Computer Science*. NS investigates non-trivial features of graph problems that usually are not addressed by lattice theory or random graphs.The understanding of such non-trivial features is of high interest, as they frequently occur in real world problems.

Our aim is to develop a common framework for knowledge management using CNA methods. Thus we can extract information from various heterogeneous domains. The development of the frameworks implies determining and adding scientific contributions to the following research fields:

1. data acquisition
2. data preprocessing
3. data storage
4. complex network creation
5. methods of analysis
6. proof of concept in various domains

In order develop and test the common framework we chose to try and resolve real world problems from the following domains: Tourism, E-learning, Freight Transportation , and Organisational Analysis. The domains just enumerated are diverse and should give

a sufficient generality to the framework to be called a common framework.

The paper is structured as follows. The next sections focuses on background information and related work. The third section briefly describes the framework. The last section presents the current status of our research and future work.

II. Background and Related Work

Two important papers stand as the building blocks of *Complex Networks Analysis*. Paul Erdös and Alfréd Rényi wrote about random graphs in 1959 [1]. In 1973, Mark Granovetter discovered the "strength of weak ties" [2]. A graph usually consists of a number of subgraphs, nodes inside these subgraphs are tightly connected among them and loosely (weak ties) connected with other subgraphs. One may think that those weak ties are not relevant, but without their presence the graph of subgraphs would not exist. CNA emerged at the beginning of the 1990's as a result of the progress in applied computational sciences. But the most important factor was the access to data describing real world networks. The emergence of the World Wide Web, as well as the explosion of the interest in detailed mapping across many sciences, especially in biology and economics, opened a multitude of research paths.

Stanley Milgram [3] and Watts et al. [4] discovered and defined the *small world phenomenon*. Otherwise called *six degrees of separation*, this phenomenon is found in many real world large networks, where contrary to the size of the network the average path length between two nodes has a very low value (6 or less). Barabasi et al. [5] showed that real world networks have a *scale free degree distribution*, also called Pareto or Zipf distribution. This means that very few nodes have high *Degree* while the majority has almost the same very low *Degree*. An explanation for the appearance of the *scale free distribution of degree* is the *preferential attachment* [6] of nodes, a node has a greater probability to be linked with nodes that have high *Degree* than with nodes with low *Degree*. Another phenomenon that is of great interest for NS is *Homophily*, described as the tendency of individuals (nodes in our case) to associate and bond with similar others [7].

CNA can be used in many application domains. For example, internet companies like *Google* and *Facebook* are practically built on complex networks. In medicine, the spread of diseases is now studied with the help of CNA [8]. Security forces map the networks of acquaintances of wanted individuals, maps which could lead to alternative ways to reach them. The famous Saddam Hussein was captured using methods from NS [9]. Large oil companies use a branch of CNA known as *Organisational Network Analysis* to enhance the flow of information exchange within the companies [10]. CNA was even used to determine the best tennis players respective to different scenarios [11], e.g. best tennis player on the grass surface.

III. Framework

The first aspect in the design of the framework should be it's universality. We are looking to develop the framework such that in can be used and easily adapted for diverse use cases no matter the domain of the problem. But we want also to put some restrictions in order to ensure the quality of the results. Therefor some of the guidelines shall be mandatory but the majority are optional. The guidelines are extracted from our experience in the already mentioned domains.

The main restriction in using the framework is modelling the domain of interest into a graph. Although this might seem a considerable restriction keep in mind that it is very easy to abstract the real work into objects and relations among the objects. By object we understand phenomenon/ living thing /material object that can be described as a sum of states at a certain point in time.

The main feature of framework is the power to analyse the resulted graph/graphs from various granularity levels:

1. from the perspective of the entire graph/network
   (a) evolution in time, with possibility to predict further evolution.
   (b) the level of resilience of the graph, with indications on how to increase or reduce the resilience.
   (c) the ability of the graph to support information/knowledge exchange between the ob-

jects, with indications on how to improve information/knowledge exchange.

(d) detection of graph particularities, with possibility of detecting similar graphs based on those particularities.

(e) social phenomenon detection, e.g. small world.

(f) knowledge extraction based on visualisation.

2. from the perspective of communities inside the graph

(a) community detection using traditional artificial intelligence algorithms, complex networks algorithms or hybrid algorithms

(b) the ability of the graph to support information/knowledge exchange between communities, with indications on how to improve information/knowledge exchange.

3. from the objects's perspective

(a) determining the objects with high centrality, with the option of developing/optimising centrality measures for particular domains.

(b) identification of particular objects.

(c) hybrid object recommendation system based on CNA metrics and other scientific methods, e.g. natural language processing.

Before each particular use of the framework the user needs to determine the objects and their defining states. Objects can be represented strictly conforming to a pattern, where the domain is well defined, or in a schema-less mode, especially useful when the domain of research is entirely regulated. E.g the tourism domain is not entirely regulated, a king size bed may also be known as a sultan size bed due to cultural differences. Relations need to be thoroughly defined.

Regarding data acquisition a multitude of tools can be used or developed depending on the on the source, e.g. web crawlers. But before a source of data is selected it is mandatory to check for its quality, garbage in garbage out. If the data is extracted from multiple sources it is mandatory to understand and consider similarities and dissimilarities between the sources in the data acquisition process, e.g. multiple definitions of the same thing need to be avoided. As much as possible include also temporal data, thus evolutionary analysis can be conducted.

Data prepossessing is not mandatory if the source of data is clean, e.g. data from U.S. patent bureau, otherwise we need to clean the data. The amount of preprocessing is research but at least duplicate, unreadable data and data that gives no added value should be eliminated. Detecting outliers and eliminating them could have a significant improvement in the end results. Natural language processing of texts can be usefull in eliminating parts of speech or stop words that represent no valuable data. Twitter tag expansion can also be valuable, as it brings relevant keywords in the analysis, e.g. from "#thebestcity" becomes "the best city". By using RDF resources like DBpedia[1] we can enrich the knowledge base.

Data can be stored in many forms and in many systems. We recommend using a database system. The choice depends on how much "joggle" with the data is needed. For very ambitious "joggle" we recommend NoSQL graph data bases, like Ne04j[2], as jumping and combining relations is very easy. If the objects that shall be analysed are schema-less and the aggregation structure needs no change then NoSQL aggregate-oriented databases are the best choice, e.g. MongoDB[3]. Otherwise traditional SQL should be used.

The creation of the complex network/networks is possibly the most important step as the way the objects are put together has significant on knowledge extraction. A "mud-ball" graph consisting all objects and all relations might give some information but usually that is not true. Thus a series of trial and-error construction of complex networks have to be attempted. A good knowledge of the research domain is needed. Usually a graph is created for each relations defined at the beginning, an only after these are analysed multigraphs[4] are created and analysed. Based on the definitions of the relations between objects the decision to create directed graphs or undirected graphs is made. We recommend

[1] http://wiki.dbpedia.org/

[2] http://neo4j.com/

[3] https://www.mongodb.org/

[4] a multigraph is a graph which is permitted to have parallel edges

using both types, as the directed graphs can better pin point objects with high centrality, while undirected graphs reveal structural objects (those objects that keep the graph together but don't have high centrality). We also recommend using weighted graphs as they are more accurate in the abstraction of a research domain.

The methods of analysis are also research domain dependent. A major part of our research focuses on developing and optimising methods / techniques / ontologies both at a general level for specific domains. Among the algorithms used by us we mention: centrality algorithms, graph topological detection algorithms (e.g. clique detection), community detection algorithms, textual complexity algorithms. Besides algorithms we also use ontologies to define states and complex networks types. We also employed statistical methods calculating correlations.

IV. Current Status and Future Work

Regarding Freight Transportation we were able to develop a system for freight brokering using ICNET negotiation algorithm and based on an ontology developed by us for an exhaustive list of freight types. Next we plan to conceive a recommender system to recommend transport companies based on their previous contracts with freight owners.

Based on touristic reviews extracted from the Internet site *AmFostAcolo.ro* we were able to analyse the graph of information exchange and extract knowledge on information exchange and network expansion. Another recommeder system is in development to suggests tourist locations based on community preferences.

Based on messages exchange by students in an e-learning environment we were able to tie the textual complexity of students to their grades. In the future we plan to conceive a grade prediction system based on students textual complexity.

On the Organisational Analysis we've proven that the SCRUM agile development method support better information exchange and innovation than the classical hierarchical scheme. Also we analysed the information exchange in a small academic organisation and we were able to identify bottlenecks and suggest improvements. For the future we plan to analyse other agile development methods.

Acknowledgment

We acknowledge support from COST Action IC1305 NETWORK FOR SUSTAINABLE ULTRASCALE COMPUTING (NESUS).

References

[1] Erdős, P., Rényi, A.: On random graphs. Publicationes Mathematicae Debrecen **6** (1959) 290–297.

[2] Granovetter, M.: The strength of weak ties. American journal of sociology **78** (1973) 1

[3] Milgram, S.: The small world problem. Psychology today **2** (1967) 60–67

[4] Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. nature **393** (1998) 440–442

[5] Barabási, A.L., et al.: Scale-free networks: a decade and beyond. science **325** (2009) 412

[6] Newman, M.E.: Clustering and preferential attachment in growing networks. Physical Review E **64** (2001) 025102

[7] McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: Homophily in social networks. Annual review of sociology (2001) 415–444

[8] Barabási, A.L., Gulbahce, N., Loscalzo, J.: Network medicine: a network-based approach to human disease. Nature Reviews Genetics **12** (2011) 56–68

[9] Wilson, C.: Searching for saddam: Why social network analysis hasn't led us to osama bin laden. Slate, February **26** (2010)

[10] Cross, R.L., Singer, J., Colella, S., Thomas, R.J., Silverstone, Y.: The organizational network fieldbook: Best practices, techniques and exercises to drive organizational innovation and performance. John Wiley & Sons

[11] Radicchi, F.: Who is the best player ever? a complex network analysis of the history of professional tennis. PloS one **6** (2011) e17249